基于支持向量机的传真收件人识别方法

林开标　王周敬
(厦门大学自动化系,福建厦门 361005)
E-mail:lkbiao208@163.com

摘　要　在字符特征提取基础上,文章提出了应用支持向量机对传真收件人进行识别的方案,解决了传真收件人格式、表示方法多样性而导致的自动分发困难的问题。文中对四种常用的核函数分别进行了实验,选取了对传真收件人具有较高识别率的核函数,它有利于实现传真文件的自动分发。

关键词　支持向量机　字符识别　传真　特征提取

文章编号 1002-8331-(2006)07-0156-03　**文献标识码** A　**中图分类号** TP391.43

The Method of Fax Receiver's Name Recognition Based on SVM

Lin Kaibiao　Wang Zhoujing
(Dept. of Automation, Xiamen University, Xiamen, Fujian 361005)

Abstract: Based on characters feature extraction, the scheme of fax receiver's recognition is advanced in this paper by the use of SVM, which solves some problems such as the format of fax receiver and the difficulty of intelligent distribution caused by the diversity of expressing methods.Four common kernel functions are experimented respectively in this paper, and the kernel function that has better recognizing accuracy to fax receiver is selected, which can make for the realization of the intelligent distribution of fax.

Keywords: Support Vector Machines(SVM), character recognition, fax, feature extraction

传真以其独有的特点在互联网盛行的今天依然拥有其原有的地位,并且凭借其技术的不断进步,得到了市场和用户的认可和喜爱,在办公自动化领域占有重要位置。传真的广泛使用,使得传真自动分发在企业中变得日益重要,实现一个传真自动分发系统,不仅便于传真统一管理和提高工作效率,还可以减少人为产生的错分可能性。但因为传真文件没有固定格式,其收件人姓名可为印刷体也可为手写体,这种使用上的灵活性给传真收件人识别带来了很多困难,使得到目前为止这一方面的研究较少。

要实现传真文件的自动分发,最主要是对收件人姓名进行识别,由于一个企业的员工人数是有限的,每个收件人姓名可能用到的称呼也有限,需要识别的字符总数不是很多,属于小样本字符集,因此,传真收件人姓名识别问题等价于小字符集样本识别问题。对此,本文将支持向量机(Support Vector Machines)引入传真收件人自动识别系统中,通过对传真收件人姓名的识别实现传真文件的自动分发。

1　支持向量机(SVM)算法

支持向量机(SVM)是AT&T Bell实验室的Vapnik等人根据统计学习理论提出的一种新机器学习方法。它的基本思想是根据Vapnik提出的结构风险最小化原理,通过最大化分类间隔或边缘尽量提高学习机的泛化性能。

支持向量机是从线性可分情况下的最优分类面发展来的。设线性可分样本集为(x_i, y_i), i=1, …, l, $x\in R^N$, $y\in\{+1, -1\}$是类别标号。N维空间中线性判别函数的一般形式为$g(x)=w\cdot x+b$,分类面方程为$w\cdot x+b=0$,该方程可以将样本无错分开,从而保证了训练错误率为0。然后对判别函数归一化,使两类所有样本都满足$|g(x)|\geq 1$,使离分类面最近样本的$g(x)=1$,分类间隔等于$2/\|w\|$。

根据统计学习理论,支持向量机应该在样本中建立最优分类面(Optimal Separating Hyperplane,简称OSH),最优分类面的一个重要特征就是使被分开的两类样本的间隔最大,因为理论分析指出,间隔最大意味着推广能力强、VC维上界最小,从而实现SRM准则中对函数复杂性的选择,这是支持向量机的核心思想之一。所以,满足下述条件:

$$y_i[(w\cdot x_i)+b]-1\geq 0,\ i=1, 2, \cdots, l \tag{1}$$

且使$\|w\|$最小的分类面就是最优分类面。利用拉格朗日优化方法把这个问题的求解转化为对偶形式,即在约束条件

$$\sum_{i=1}^{l} y_i a_i=0 \quad a_i\geq 0,\ i=1, \cdots, l \tag{2}$$

下求解下列函数的最大值:

$$W(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j) \tag{3}$$

这个方程存在唯一解,其中只有少部分α_i不为零,所对应的样本就是离最优分类面最近的支持向量。得到的最优分类函数是:

$$f(x)=\mathrm{sgn}\{\sum_{i=1}^{l} y_i\alpha_i x_i\cdot x+b\} \tag{4}$$

当遇到线性不可分样本的时候,可以通过非线性变换$\Phi: R^N\rightarrow F$把样本映射到某个高维空间F(称作特征空间)里,在那里进行线性分类,这样分类方程变为:

作者简介:林开标,男,硕士研究生,研究方向:模式识别与图像处理。王周敬,男,副教授,研究方向:模式识别与图像处理、人工智能、数据挖掘。

$$f(x)=\mathrm{sgn}\{\sum_{i=1}^{l} y_i\alpha_i\Phi(x_i)\cdot\Phi(x)+b\} \quad (5)$$

根据泛函理论，只要一种内积函数 $K(x,y)$ 满足 Mercer 条件，它就对应某一变换空间的内积，因此用内积函数代替内积得到最终的分类函数，也就是支持向量机：

$$f(x)=\mathrm{sgn}\{\sum_{i=1}^{l} y_i\alpha_i K(x_i,x)+b\} \quad (6)$$

在这里，约束条件变为：

$$0\leq\alpha_i\leq C, i=1,\cdots,l \quad (7)$$

其中，C 是约束参数，可根据具体情况确定。因此，通过选择合适的内积函数，就可以实现某一非线性变换后的线性可分。

2 传真收件人识别方法

通过对大量传真样本的统计分析，得到传真收件人的识别具有以下特点：

(1) 一个企业的员工人数是有限的，每个收件人姓名可能用到的称呼也有限，因此，需要识别的字符总数不是很多，属于小样本字符集；

(2) 收件人姓名的前方基本上都有标识符，如“收件人：”、“To:”、“ATTN:”等，可通过匹配这些关键字来定位收件人姓名部分的图像；

(3) 收件人姓名位置很可能在表格中或下划线上，根据这个特征，结合第(2)个特征，可较快且更准确地定位到收件人姓名部分的图像；

(4) 收件人姓名可能是印刷体，也可能是手写体，在字符识别之前必须让程序先判别；

(5) 收件人姓名中可能包含有汉字、英文、数字等字符，在字符切分时必须充分考虑各种可能情况；

(6) 收件人姓名中不同字符的出现频率差异较大，如“先”、“生”、“小”、“姐”等反复出现，而有些字符只是偶尔出现。

2.1 传真收件人识别系统的组成

结合以上一些特点，传真收件人识别系统共由三大模块组成：图像预处理模块、版面分析模块、文字识别模块。

由于受到种种条件的限制和随机干扰，传真原始图像中包含有各种各样的噪声和畸变，为了版面分析、字符切分等后续操纵的需要，要对取得的图像作预处理，其处理效果将直接影响后续处理方法的难易及结果的准确性，图像预处理模块的主要内容是灰度校正、噪声过滤、二值化和倾斜校正。版面分析指的是对扫描得到的图像进行分割，将图像中可能包含的文本、图形、表格、图像、标题等版面基元区分开，并得到它们的逻辑关系。传真收件人识别的版面分析的功能是将收件人姓名部分的图像识别并截取出来，该模块在关键字的特征匹配时需调用文字识别模块的一些功能，即先识别出如“收件人：”、“To:”、“ATTN:”等标识收件人姓名位置的特殊字符，然后定位并截取收件人姓名部分的图像，由于篇幅有限，这部分的内容将在另文介绍。本文重点介绍文字识别模块，该模块包括对字符图像做预处理、特征提取和分类识别等操作。

传真收件人识别系统实现了传真文件收件人姓名的识别，但要实现传真自动分发还必须包括智能分发模块。该模块将已提取的收件人姓名与数据库中的用户进行匹配，匹配成功，将该传真发送到收件人的内部邮箱中或以某种方式提醒收件人，匹配不成功，转为人工发送。智能分发模块除了具有自动分类、发送功能外，还具有自学习功能，即在匹配不成功时，经过人工发送，系统能够自学习，在下次遇到同类传真时，可以自动识别。

2.2 字符图像的预处理

字符图像预处理在功能上不同于传真收件人识别系统的图像预处理模块，它是对待识别字符图像做预处理，而不是对整张传真图像做预处理。

字符图像预处理是文字识别的第一个阶段，我们首先定位要识别的收件人姓名字符；由于纸张的厚薄、洁白度、光洁度、油墨深浅、书写质量等都要造成字形畸变，产生污点、飞白等干扰，因此采用 Urger 平滑法对字符图像进行去噪；再用直方图投影分割出字符，归一化到 24×24 像素点阵。因为印刷体字符和手写体字符的预处理过程相似，所以本文只给出手写体字符的处理结果，如图 1 所示。

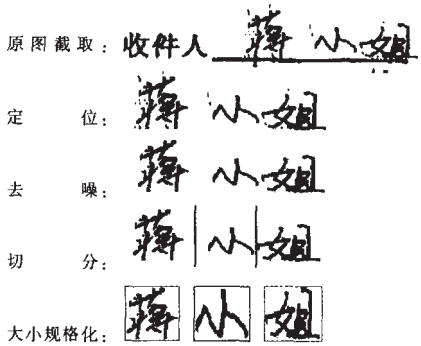

图 1 手写体处理效果

2.3 特征提取

特征提取是汉字识别的一个重要环节，它是从单个字符图像上提取统计特征或结构特征的过程，因此也可以看作是一个降维的过程，所提取特征的好坏将直接影响整个汉字识别系统的性能。根据原始点阵直接求出的统计特征通常反映图像点阵分布的总体情况，这类特征的图像预处理简单、对噪声不敏感，但对一些精细部分的反应不灵敏；而由骨架和轮廓得到的结构分类方法，对于精细部分结构反应灵敏，但对噪声敏感。采用单一的特征提取方法利用的汉字信息量有限，不可避免地会存在一些识别的“死角”，也就是存在利用该特征很难区分的汉字。因此，传真收件人识别系统将统计特征和结构特征相结合，扬长避短，发挥各自优点，提取了外围轮廓特征、投影特征、网格点阵特征、Hu 矩不变量特征等共二百多维的特征。

以外围轮廓特征为例，其特征提取过程如下：针对规格为 24×24 的二值图像字符样本，按先后顺序从左、右、上、下四边分别向右、左、下、上四个方向扫描，直至扫描线遇到字符像素点或与扫描线垂直的中轴，记下各自扫描线走过的距离，即为该字符的外围轮廓特征，是一个 24×4=96 维的特征。

2.4 分类识别方法

由于传真收件人姓名中字符出现频率差别较大，如“先”、“生”、“小”、“姐”等字符反复出现，而有些姓名字符很少出现。所以为了提高训练系统的识别率和运行效率，本文首先判别待识别字符是印刷体还是手写体，然后根据字符出现的频率不同对字符采取分级识别策略：将字符样本分为 3 级，其中出现频率最高的为第 1 级，偶尔才会出现的字符为第 3 级，其余的为

第 2 级。对待识别字符，先用第 1 级字符样本进行判别，能识别则结束；不能识别，改用第 2 级样本进行识别；依此类推，直到第 3 级样本还不能识别，则判别该字符为拒识字符。

对于每一级字符样本，本文采用多个二分类器组合的一对多(One Versus Rest, OVR)算法，将多类识别问题转化为二类识别问题来解决。每个分类器只将一个汉字与其余汉字区分开，训练样本中该汉字对应的 y 值为+1，其余样本对应的 y 值为-1。首先确定使用的核函数 K，将训练 k 样本值带入优化函数中，求出最优解及其非零值对应的支持向量，并根据任一训练样本值求出阈值 b。即可求出所有参数值，得到判别函数 f(x)。依此类推，分别求出所有汉字对应的判别函数。进行字符识别时将输入信号送到每一个分类器，然后循环检查所有的分类器输出。若某一分类器的输出值为“1”，则认为输入的字符为该类对应的汉字字符；否则，若所有输出值均不为“1”，则拒绝识别该字符。由于识别存在一定的误差，可能同时有多个分类器的输出值为“1”，此时则判断为第一个输出值为“1”的字符类。

2.5 实验

本文的实验样本取自某企业 2004 年度下半年共 1 028 封传真，经统计分析，其中有 912 封只属于 14 个人，为便于取样，我们只对这 14 个人所用到的姓名及称呼字符进行实验，其中有 205 封传真的收件人字符是手写体，共用到 22 个汉字，无英文字母，其余为印刷体，共用到 65 个字符，其中 12 个是英文字母。实验过程中，我们选取印刷体字符样本 50 套，30 套作为训练集，20 套作为测试集，选取手写体字符 30 套，20 套作为训练集，10 套作为测试集，由于有些字符出现次数未达到取样数，我们采取在字符样本上加上不同的噪声处理来进行样本集的扩充，取惩罚因子 C 等于 100，采用四种常用的核函数的 SVM 算法先对印刷体字符的识别进行实验，所得测试结果如表 1～表 4 所示。

表 1 线性核函数 $K(x,y)=x\cdot y$

平均 SV 数	平均消耗时间/s	平均识别率/%
18	0.403	97.22

表 2 多项式内积函数 $K(x,y)=[(x\cdot y)/256]^d$

d	平均 SV 数	平均消耗时间/s	平均识别率/%
1	17	0.428	98.20
2	19	0.442	98.11
3	21	0.435	97.64
4	21	0.423	97.38
5	20	0.432	97.21
6	18	0.422	97.12

表 3 径向基内积函数(RBF) $K(x,y)=\exp\left|-\frac{|x-y|^2}{256\sigma^2}\right|$

σ^2	平均 SV 数	平均消耗时间/s	平均识别率/%
0.1	16	0.434	98.42
0.2	15	0.425	98.55
0.5	15	0.433	99.10
1.0	17	0.424	98.92
1.2	17	0.428	98.60
2.0	19	0.440	98.64

印刷体字符的识别结果表明，应用支持向量机对传真收件人字符进行识别，能在较短的时间内获得较高的识别率。同时发现，不同核函数的识别时间差别不是很大，但对识别率有影响，其中径向基内积函数(RBF)的平均识别率比其它的核函数的识别率要高些。所以本文系统选取径向基内积函数(RBF)作为支持向量机的核函数。该函数对手写体字符的测试结果如表 5 所示。

表 4 神经网络内积函数 $K(x,y)=\tanh(a(x\cdot y)/256+b)$

a	b	平均 SV 数	平均消耗时间/s	平均识别率/%
1	0.8	17	0.440	97.15
	0.9	17	0.429	97.22
	1.0	18	0.430	96.83
	1.1	19	0.435	97.14
2	0.8	16	0.428	97.35
	0.9	16	0.429	97.61
	1.0	17	0.431	97.20
	1.1	17	0.435	97.14

表 5 手写体字符识别结果

σ^2	平均 SV 数	平均消耗时间/s	平均识别率/%
0.1	11	0.422	92.62
0.2	11	0.423	92.42
0.5	10	0.418	94.08
1.0	12	0.431	93.52
1.2	12	0.426	92.40
2.0	11	0.412	92.64

径向基内积函数(RBF)作为核函数对手写体字符的测试结果表明，手写体字符识别率比印刷体字符的识别率要低一些，其原因可以归结为汉字规模大、相似汉字较多且手写体汉字存在大量不规则变形。同时发现，在核函数的参数变化时，识别率会略有不同，如在 σ^2 取 0.5 时字符的识别率较高，因此，在实际使用时，必须选择合适的参数，以提高系统的识别率和识别速度。本文系统选取 σ^2 为 0.5 作为径向基内积函数(RBF)的参数值。

3 结束语

本文在字符特征提取基础上，应用支持向量机对传真收件人进行识别，解决了因为传真收件人格式、表示方法多样性而导致的自动分发困难的问题。实验结果表明，采用支持向量机能够解决传真收件人姓名的识别问题。

由于收件人的手写体字符识别率比印刷体字符的识别率要低的多(5.02%)，因此如何提高手写体字符的识别率是下一步需要研究的问题。(收稿日期：2005 年 9 月)

参考文献

1.(英)Nello Cristianini, John Shawe-Taylor 著.李国正，王猛，曾华军译.支持向量机导论[M].北京：电子工业出版社，2004

2.Vapnik V.张学工译.统计学习理论的本质[M].北京：清华大学出版社，1999

3.Platt J C.Sequential minimal optimization: a fast algorithm for training support vector machines.Advances in kernel methods support vector learning[M].Cambridge, MA: MIT Press, 1999: 185～208

4.高彦宇，杨扬.脱机手写体汉字识别研究综述[J].计算机工程与应用，2004; 40(7): 74～77

5.Schomaker L, Vuurpijl L.Support Vector Machines for the classification of western handwritten capitals[C].In: Proceeding of Seventh International Workshop on Frontiers in Handwriting Recognition, 2000: 167～176